Are Guidelines Following Guidelines?

The Methodological Quality of Clinical Practice Guidelines in the Peer-Reviewed Medical Literature

Terrence M. Shaneyfelt, MD, MPH
Michael F. Mayo-Smith, MD, MPH
Johann Rothwangl, MD, FACG

Context Practice guidelines play an important role in medicine. Methodological principles have been formulated to guide their development.

Objective To determine whether practice guidelines in peer-reviewed medical literature adhered to established methodological standards for practice guidelines.

Design Structured review of guidelines published from 1985 through June 1997 identified by a MEDLINE search.

Main Outcome Measures Mean number of standards met based on a 25-item instrument and frequency of adherence.

Results We evaluated 279 guidelines, published from 1985 through June 1997, produced by 69 different developers. Mean overall adherence to standards by each guideline was 43.1% (10.77/25). Mean (SD) adherence to methodological standards on guideline development and format was 51.1% (25.3%); on identification and summary of evidence, 33.6% (29.9%); and on the formulation of recommendations, 46% (45%). Mean adherence to standards by each guideline improved from 36.9% (9.2/25) in 1985 to 50.4% (12.6/25) in 1997 ($P<.001$). However, there was little improvement over time in adherence to standards on identification and summary of evidence from 34.6% prior to 1990 to 36.1% after 1995 ($P=.11$). There was no difference in the mean number of standards satisfied by guidelines produced by subspecialty medical societies, general medical societies, or government agencies ($P=.55$). Guideline length was positively correlated with adherence to methodological standards ($P=.001$).

Conclusion Guidelines published in the peer-reviewed medical literature during the past decade do not adhere well to established methodological standards. While all areas of guideline development need improvement, greatest improvement is needed in the identification, evaluation, and synthesis of the scientific evidence.

JAMA. 1999;281:1900-1905 www.jama.com

CLINICAL PRACTICE GUIDELINES are commonly defined as "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances."[1] Over the past decade there has been a surge of interest in the use of clinical practice guidelines fueled by the discovery of large, unexplained variation in physician practice,[2-6] documentation of significant rates of inappropriate care,[7] and an interest in managing health care costs.[8] It is believed that practice guidelines can improve the quality, appropriateness, and cost-effectiveness of health care,[1] and can also serve as valuable educational tools.[9]

In response to this increased interest, several major medical organizations, including the American Medical Association (AMA), the Institute of Medicine (IOM), and the Canadian Medical Association, have carefully formulated methodology for developing scientifically sound guidelines.[1,10-13]

The purpose of this study was to systematically examine guidelines published in the peer-reviewed medical literature to determine to what degree they use and document these methodological standards, in which areas they may be deficient, and whether there were changes over time.

For editorial comment see p 1950.

METHODS

Instrument Development

Using the principles formulated by the major medical organizations mentioned, a group of experts in guidelines and evidence-based medicine identified key elements for the development and reporting of guidelines.[14]

Author Affiliations: Division of General Internal Medicine, Department of Medicine, Beth Israel Deaconess Medical Center, and Department of Medicine, Harvard Medical School, Boston, Mass (Drs Shaneyfelt and Mayo-Smith); and the Ambulatory Care Service, Veterans Affairs Medical Center, Manchester, NH (Drs Shaneyfelt, Mayo-Smith, and Rothwangl). Dr Shaneyfelt is now with the Division of General Internal Medicine, University of Alabama at Birmingham.
Corresponding Author and Reprints: Terrence M. Shaneyfelt, MD, MPH, The University of Alabama at Birmingham, Division of General Internal Medicine, 621 Medical Education Bldg, 1813 Sixth Ave S, Birmingham, AL 35294-3296 (e-mail: tshaneyfelt@gim.dom.uab.edu).

The elements were formulated by this group through a careful series of review and pilot testing by guideline developers, evaluators, implementers, and groups of practicing clinicians. Reviewers included consultants at the National Library of Medicine, the IOM, the American College of Physicians, and the Agency for Health Care Policy and Research. Feedback was also solicited at 3 national workshops about practice guidelines. Based on the careful, comprehensive, and inclusive development process used, we felt that these criteria were a valid representation of current standards for guidelines. We developed a 25-item instrument, using a yes or no format, to measure adherence to these elements, broadly grouped into standards on guideline format and development (10 items), identification and summary of evidence (10 items), and formulation of recommendations (5 items). Questions were refined for clarity through multiple rounds of pretesting by the authors on 35 published guidelines.

To further confirm content validity, we surveyed 13 experts who have published articles on guideline methodology and a random sample of persons responsible for guideline development for 12 major medical organizations to independently evaluate the validity of our instrument as a measure of the methodological quality of guidelines. The median rating of validity on a scale of 1 to 5 (1 representing poor validity and 5 excellent validity) by both the experts and developers was a 4 (range, 3.5-5.0) with 92% rating it 4 or 5. Finally, to determine if the instrument could differentiate well-developed from poorly developed guidelines, we asked 6 persons formally trained in critical appraisal skills to rank 6 guidelines in order of quality based on accepted principles of guideline development.[1,11,12,15] In all instances, the instrument ranked the guidelines in the same order as the reviewers. Thus, overall, we feel our instrument is a valid measure of the methodological quality of guidelines.

Guideline Selection and Evaluation

Guidelines were identified by a computerized search of the MEDLINE database from January 1966 through June 1997 using the following terms: *practice guideline (pt)*, *guideline*, *practice parameters*, *protocols*, *consensus conferences or statements*, *algorithms*, *standards*, and *practice policies*. Names of medical organizations and government agencies involved in practice guideline activity were also included as search terms. The 1996 *AMA Directory of Practice Parameters*,[16] as well as the bibliographies of guidelines, editorials, review articles, and other articles about guidelines, were searched for additional published guidelines.

Retrieved documents were considered guidelines if they met the definition of a guideline as proposed by the IOM.[1] We excluded articles on diagnostic criteria or technical standards, guidelines on research methods, review articles, and any secondary publications of the guideline. Since few guidelines were published prior to 1985 and because of the large number of guidelines published overall, we evaluated only guidelines published in peer-reviewed journals in odd-numbered years from 1985 through June 30, 1997. In addition, we retrieved and evaluated any background supporting articles cited as part of the guideline if available in peer-reviewed journals.

Each guideline was independently evaluated by 2 investigators for adherence to methodological standards. The level of agreement for independent reviews was 87% ($\kappa = 0.73$, a rate of agreement considered to be substantial[17]). Discrepancies were resolved by open discussion and, in less than 10% of the guidelines, by adjudication of a third reviewer blinded to the previous reviews.

Statistical Analysis

The total number of standards satisfied by each individual guideline could range from 0 to 25. The mean (SD) number of standards satisfied was calculated collectively and for each individual year. In addition, the frequency of adherence to each of the 25 standards was calculated. We evaluated time trends of adherence to methodological standards by constructing multiple crude linear regression models with "year" as the independent variable (1 *df*).

Guideline developers were divided into 4 groups: general and subspecialty medical societies, government agencies, and others that included individuals, insurers, and private organizations. The mean number of standards satisfied by guidelines produced by each of these groups was compared by 1-way analysis of variance.

Finally, we tested for the effect of prior experience producing guidelines and length of the guideline document on adherence to methodological standards by constructing multiple linear regression models. In these models, the independent variable was used in a continuous fashion with 1 *df*.

The level of statistical significance was established at a 2-sided *P* value of less than .05. (All analyses were performed using the *Statistical Analysis System*, version 6.12 [1997], SAS Institute Inc, Cary, NC.)

RESULTS

We evaluated 279 guidelines covering a wide range of topics. (A bibliographic list of evaluated guidelines is available from the authors on request.) Overall, the mean (SD) number of standards satisfied (out of 25) was 10.77 (3.71), or 43.1%, with a range of 2 to 24 (FIGURE 1). Guidelines did show significant improvement from 1985 (9.2/25 or 36.9%) to 1997 (FIGURE 2), but still only 50.4% (12.6/25) of the standards were met, on average, for each guideline in 1997.

Mean overall adherence to methodological standards on guideline development and format was only fair (51%), as shown in TABLE 1. Even though guidelines are developed to improve health outcomes, only 40% specified the outcomes of interest. Fewer than half described the patient population to which the guideline applied, while slightly greater than half described the

Figure 1. Distribution of the Mean Number of Methodological Standards Satisfied by Guidelines

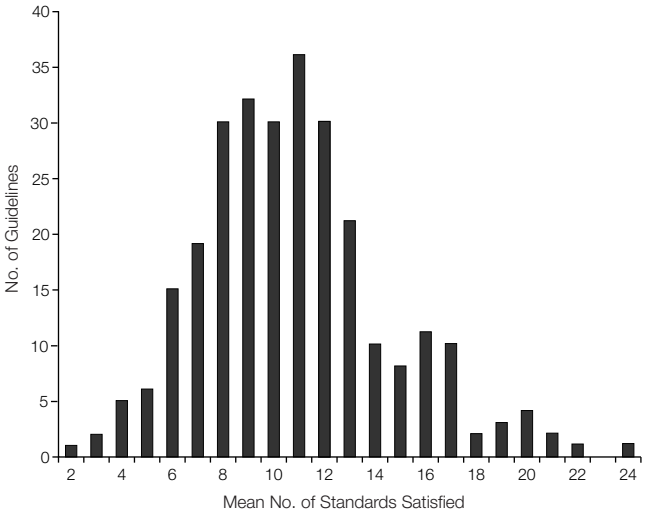
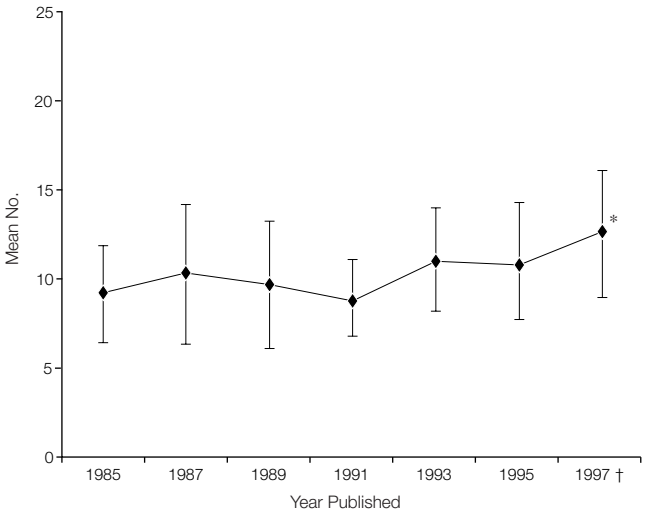


Figure 2. Mean Number of Methodological Standards Satisfied by Guidelines



Asterisk indicates $P<.01$ for trend across all years; dagger indicates that only guidelines published through June 1997 are included. Error bars indicate SDs.

Table 1. Frequency of Adherence to Methodological Standards on Guideline Development and Format

Standard	No. (%) of Guidelines Satisfying Standard (N = 279)
1. Purpose of the guideline is specified	210 (75.3)
2. Rationale and importance of the guideline are explained	244 (87.5)
3. The participants in the guideline development process and their areas of expertise are specified	72 (25.8)
4. Targeted health problem or technology is clearly defined	170 (60.9)
5. Targeted patient population is specified	128 (45.9)
6. Intended audience or users of the guideline are specified	142 (50.9)
7. The principal preventive, diagnostic, or therapeutic options available to clinicians and patients are specified	229 (82.1)
8. The health outcomes are specified	111 (39.8)
9. The method by which the guideline underwent external review is specified	90 (32.3)
10. An expiration date or date of scheduled review is specified	30 (10.8)
Mean (SD) overall adherence, %	51.1 (25.3)

intended audience of the guideline. Most guidelines (82.1%), however, specified the preventive, diagnostic, or therapeutic options available to clinicians and patients.

The methodological standards on the identification and summary of evidence were poorly adhered to, with an overall mean adherence of 33.6% (TABLE 2). Few guidelines specified the methods used to identify scientific evidence (16.8%) or the time period from which the evidence was collected (14.3%). Surprisingly few guidelines (7.5%) reported or used formal methods (eg, meta-analysis) to combine scientific data or, when data were lacking, formal methods of determining expert opinion (eg, the Delphi method). Even though guidelines have been championed as a means to decrease health care expenditures, only 41.6% made any mention of projected effects on health care costs, and only 14.3% quantified these estimates in any way. Guidelines did better in specifying the benefits and harms expected to result from specific health practices (86.4%), but only 60.2% quantified the magnitude of the benefits and harms.

Similarly, guidelines adhered poorly to methodological standards on the formulation of recommendations, with overall compliance of 46% (TABLE 3). Only 6.1% of the guidelines discussed the values used by the developers to judge the desirability of alternative practices and outcomes and to make recommendations. Moreover, few guidelines (21.5%) discussed the role of patient preferences in choosing among available options. However, all guidelines made specific recommendations for practice, and most (89.6%) discussed flexibility of the recommendations.

Guidelines significantly improved in their mean adherence to standards on guideline development and format, from 41.5% prior to 1990 to 55.9% after 1995 ($P<.001$). However, the mean adherence to standards on evidence evaluation changed little, from 34.6% prior to 1990 to 36.1% after 1995 ($P=.11$), while adherence to standards on the formulation of recommendations modestly improved from 42.8% prior to 1990 to 48.4% after 1995 ($P=.003$) (FIGURE 3).

Of the guidelines reviewed, 45% were produced by subspecialty medical societies, 33% by general medical societies, 16% by government agencies, and 6% by miscellaneous groups that could not be classified into any of the previ-

ous categories. The mean number of standards satisfied by guidelines produced by the 3 major groups did not differ significantly (10.47 vs 10.93 vs 10.32, respectively; $P = .55$).

The median number of guidelines identified for each of the 69 guideline developers was 9 (interquartile range, 3-25). Guidelines produced by organizations that developed more guidelines tended to adhere to fewer methodological standards, although this trend did not reach statistical significance ($P = .15$).

The mean (SD) length of the guidelines in our study was 9.54 (10.65) pages (range, 1-96 pages). We found that guideline length was positively correlated to adherence to methodological standards ($P = .001$). For example, the mean number of standards satisfied by guidelines fewer than 4 pages long was 7.73, while that of guidelines more than 10 pages long was 13.52.

COMMENT

The issues underlying the increasing interest in clinical practice guidelines—concerns over quality of care, marked variation in physician practice, and increasing costs—are major challenges to the medical profession. Many believe that guidelines could be a vehicle to address these problems, and it was in this belief that leading medical organizations formulated standards to guide the development of guidelines. Our findings, however, demonstrate that, to date, published guidelines are falling considerably short of these standards and that much more attention is needed by those involved in both guideline creation and in guideline review and publication.

Specific improvements are needed in several areas. The first set of criteria we studied, standards on guideline development and format, involve simply making explicit various elements of guideline purpose and content—clearly defining the health problem or technology, patient population, targeted users, outcomes, and the like. Such statements are analogous to defining the patient population, interventions, and outcomes of interest in clinical studies. For guidelines, these statements are intended to communicate clearly elements critical to the appropriate use of the guideline. Furthermore, many of these standards can be met easily and briefly without adding greatly to the length of the document.

Unfortunately, guidelines are most deficient in the identification and sum-

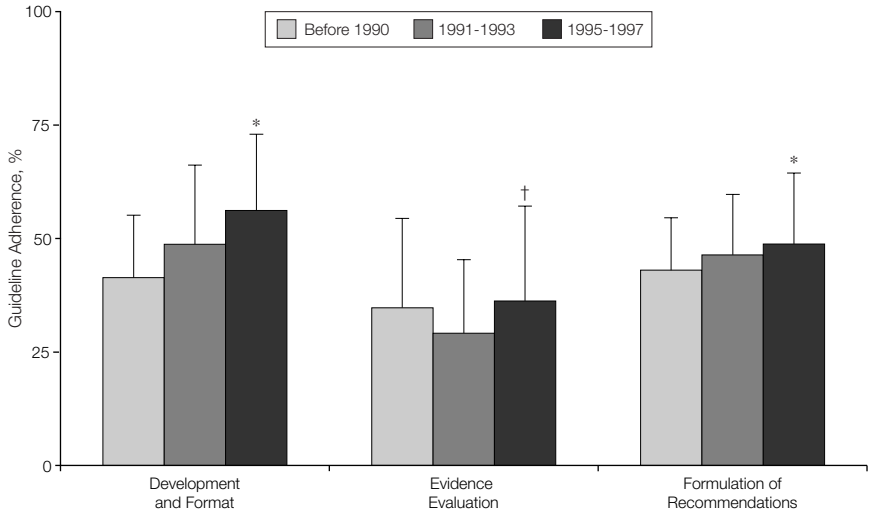
Table 2. Frequency of Adherence to Methodological Standards on Evidence Identification and Summary

Standard	No. (%) of Guidelines Satisfying Standard (N = 279)
11. Method of identifying scientific evidence is specified	47 (16.8)
12. Time period from which evidence is reviewed is specified	40 (14.3)
13. The evidence used is identified by citation and referenced	207 (74.2)
14. Method of data extraction is specified	14 (5)
15. Method for grading or classifying the scientific evidence is specified	43 (15.4)
16. Formal methods of combining evidence or expert opinion are used and described	21 (7.5)
17. Benefits and harms of specific health practices are specified	241 (86.4)
18. Benefits and harms are quantified	168 (60.2)
19. The effect on health care costs from specific health practices is specified	116 (41.6)
20. Costs are quantified	40 (14.3)
Mean (SD) overall adherence, %	33.6 (29.9)

Table 3. Frequency of Adherence to Methodological Standards on the Formulation of Recommendations

Standard	No. (%) of Guidelines Satisfying Standard (N = 279)
21. The role of value judgments used by the guideline developers in making recommendations is discussed	17.6 (6.1)
22. The role of patient preferences is discussed	60 (21.5)
23. Recommendations are specific and apply to the stated goals of the guideline	279 (100)
24. Recommendations are graded according to the strength of the evidence	36 (12.9)
25. Flexibility in the recommendations is specified	250 (89.6)
Mean (SD) overall adherence, %	46 (45)

Figure 3. Adherence to Methodological Standards Over Time



Asterisk indicates $P<.01$ for trend across all years; dagger indicates $P = .11$ for trend across all years. Error bars indicate SDs.

mary of evidence. A properly performed evaluation of the scientific evidence is critical in ensuring the scientific validity of a guideline. Less than 10% of the guidelines used and described formal methods of combining scientific evidence or expert opinion. Many used informal techniques such as narrative summaries prepared by clinical experts, a type of review shown to be of low mean scientific quality and reproducibility.[18] Indeed, it was difficult to determine if some of the guidelines made any attempt to review evidence, as less than 20% specified how evidence was identified, and more than 25% did not even cite any references. The appropriate performance of systematic reviews has been well described,[19-22] and these approaches need to be better incorporated into the formulation of guidelines.

An important goal for guidelines is to increase the efficiency in use of health care resources. Almost 60% of guidelines, however, did not mention costs at all, and only 14% provided any quantitative cost estimates. Clearly, if guidelines are to improve the cost-effectiveness of health care, greater attention must be given to economic analysis.

The third group of standards relate to the formulation of recommendations. We were able to identify specific recommendations for clinical practice in all guidelines reviewed, and overall the guidelines did well in the area of flexibility, with 89.6% specifying patient or practice characteristics justifying individualization or departure from the recommendations. Few guidelines (21.5%), however, discussed the role of patient preferences in choosing among the various health care options. Given the increasing appreciation of the importance of patient values in many clinical decisions, we believe this factor has not been adequately addressed in guidelines to date.[23] Finally, very few guidelines (6%) described the values used by the developers in judging the desirability of the various outcomes and making recommendations, leading us to believe that the importance of developer values in guidelines is not yet fully appreciated. Guidelines and systematic reviews are substantially different.[24] A guideline must not only review evidence, but must also weigh various outcomes—positive and negative—and make recommendations. The value of the various outcomes may differ significantly depending on one's perspective, and such differences may explain differences in recommendations that have occurred. For example, an organization dedicated to reducing harm from cancer may place greater value on selected cancer screening interventions, even though such interventions might prove to be extremely costly for the magnitude of the benefit they provide. Another organization, whose purpose is to promote the overall health of society, may view the same evidence differently, preferring to concentrate on other proven interventions with greater impact on overall public health. Examples of this are the conflicting recommendations among current breast cancer and prostate cancer screening guidelines.[25-28] Clearly, guideline developers need to give more attention to this aspect of guideline development, reflecting on the values they hold in terms of how they weigh evidence and view the importance of specific outcomes.

There are several limitations to our study. First, we did not differentially weigh the relative importance of the individual standards, though some may have a more central role in creating a scientifically valid, clinically useful guideline. Nevertheless, the standards we reviewed have been identified by expert groups as being important. We did not want to create a "quality index" for comparing individual guidelines, but rather we wanted to assess the quality of the guideline literature as a whole in meeting standards already developed by expert groups.

Second, by using a "yes/no" format we could not assess the relative quality of a guideline's compliance with a given standard. For example, the criterion on quantification of costs was met with great variation, ranging from the presentation of the cost of 1 drug to sophisticated economic analyses. We gave the guidelines credit, however, if any quantitative cost information was present, even though for many it was very limited. This approach of holding guidelines to only the simplest criteria was used for all standards and makes the poor performance of the reviewed guidelines of more concern, for had they been held to more detailed criteria, they would have done even worse.

Third, because we relied on material reported in the published versions of the guidelines, our findings could be affected not only by the quality of the guidelines themselves, but also by the quality of the reporting process. It is possible that in some cases guideline developers used appropriate techniques but did not report them. We attempted to minimize this by including in our evaluation any background supporting articles cited as part of the guideline if they were available in the peer-reviewed literature. However, we also feel that just as in other types of medical reports, documentation of methods used is important, as the validity of the recommendations can only be determined if the methods used to develop them are explicitly stated.

Finally, our review was limited to guidelines published in the peer-reviewed medical literature. We felt such guidelines would be most readily identifiable and available to clinicians. Furthermore, we felt because of the peer review process, such guidelines could most legitimately be held accountable to methodological criteria.

Several approaches could be used to improve the quality of guidelines. First, guideline producers could become more familiar with guideline development standards and make greater efforts to incorporate them into guidelines. A more formal effort could also be made by journal editors to require certain criteria be met prior to consideration for publication, as is currently being done for randomized controlled trials[29] or meta-analyses.[30] The AMA has proposed a Clinical Practice Guideline Recognition Program to recognize those guidelines meeting certain standards. In addition, the Agency for

Health Care Policy and Research has developed a National Guideline Clearinghouse, a Web site containing full-text guidelines along with standardized information on the methodology used to develop the guidelines.[31] These are laudable efforts to recognize the need for guidelines to meet certain standards and they reward those that do.

Guidelines, at least in the foreseeable future, will continue to be developed as tools to improve the quality of patient care. Standards for guideline development have been established, and while we anticipate their refinement over the next several years, they represent the current "state of the art" and guideline developers should strive to widely adopt and use them. This would be an important step to helping guidelines live up to their potential as a means of improving patient care and health outcomes.

Acknowledgment: The authors wish to thank Russell Phillips, MD, for his insightful review of the manuscript and Roger Davis, ScD, for his statistical consultation. We are indebted to all those who reviewed our instrument during its development.

REFERENCES

1. Committee to Advise the Public Health Service on Clinical Practice Guidelines, Institute of Medicine. Field MJ, Lohr KN, eds. *Clinical Practice Guidelines: Directions of a New Program.* Washington, DC: National Academy Press; 1990.
2. Lewis CE. Variations in the incidence of surgery. *N Engl J Med.* 1969;281:880-884.
3. Wennberg J, Gittelsohn A. Small-area variations in health care delivery. *Science.* 1973;182:1102-1108.
4. Stockwell H, Vayada E. Variations in surgery in Ontario. *Med Care.* 1979;17:390-396.
5. Roos NP. Hysterectomy: variations in raters across small areas and across physicians' practices. *Am J Public Health.* 1984;74:327-335.
6. Chassin MR, Brook RH, Park RE, et al. Variations in the use of medical and surgical services by the Medicare population. *N Engl J Med.* 1986;314:285-290.
7. Chassin MR, Kosecoff J, Park RE, et al. Does inappropriate use explain geographic variations in the use of health care services? a study of three procedures. *JAMA.* 1987;258:2533-2537.
8. Woolf SH. Practice guidelines: a new reality in medicine, I: recent developments. *Arch Intern Med.* 1990; 150:1811-1818.
9. Berg AO, Atkins D, Tierney W. Clinical practice guidelines in practice and education. *J Gen Intern Med.* 1997;12(suppl 2):S25-S33.
10. Eddy DM. *A Manual for Assessing Health Practices and Designing Practice Policies: The Explicit Approach.* Philadelphia, Pa: American College of Physicians; 1992.
11. American Medical Association, Office of Quality Assurance. *Attributes to Guide the Development and Evaluation of Practice Parameters.* Chicago, Ill: American Medical Association; 1990.
12. Canadian Medical Association. *Quality of Care Program: The Guidelines for Canadian Clinical Practice Guidelines.* Ottawa, Ontario: Canadian Medical Association; 1993.
13. Woolf SH. *Manual for Clinical Practice Guideline Development.* Rockville, Md: Agency for Health Care Policy and Research; 1991. AHCPR publication 91-0007.
14. Hayward RSA, Wilson MC, Tunis SR, Bass EB, Rubin HR, Haynes RB. More informative abstracts of articles describing clinical practice guidelines. *Ann Intern Med.* 1993;118:731-737.
15. Hayward RS, Wilson MC, Tunis SR, Bass EB, Guyatt GH. Users' guides to the medical literature, VIII: how to use clinical practice guidelines, A: are the recommendations valid? *JAMA.* 1995;274:570-574.
16. *Directory of Practice Parameters-1995 Edition.* Chicago, Ill: American Medical Association; 1995.
17. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33:159-174.
18. Oxman AD, Guyatt GH. The science of reviewing research. *Ann N Y Acad Sci.* 1993;703:125-134.
19. Chalmers I, Altman DG, eds. *Systematic Reviews.* London, England: BMJ Publishing Group; 1995.
20. Woolf SH. *Manual for Conducting Systematic Reviews.* Rockville, Md: Agency for Health Care Policy and Research; 1996.
21. Counsell C. Formulating questions and locating primary studies for inclusion in systematic reviews. *Ann Intern Med.* 1997;127:380-387.
22. Meade MO, Richardson SW. Selecting and appraising studies for a systematic review. *Ann Intern Med.* 1997;127:531-537.
23. Hlatky MA. Patient preferences and clinical guidelines. *JAMA.* 1995;273:1219-1220.
24. Cook DJ, Greengold NL, Ellrodt AG, Weingarten SR. The relation between systematic reviews and practice guidelines. *Ann Intern Med.* 1997;127:210-216.
25. Leitch AM, Dodd GD, Costanza M, et al. American Cancer Society guidelines for the early detection of breast cancer: update 1997. *CA Cancer J Clin.* 1997; 47:150-153.
26. National Institutes of Health Consensus Development Panel. National Institutes of Health Consensus Development Conference Statement: breast cancer screening for women ages 40-49, January 21-23, 1997. *J Natl Cancer Inst.* 1997;89:1015-1020.
27. US Preventive Services Task Force. *Guide to Clinical Preventive Services.* 2nd ed. Baltimore, Md: Williams & Wilkins; 1996.
28. American College of Physicians. Screening for prostate cancer. *Ann Intern Med.* 1997;126:480-484.
29. Begg C, Cho M, Eastwood S, et al. Improving the quality of reporting of randomized controlled trials: the CONSORT statement. *JAMA.* 1996;276:637-639.
30. Cook DJ, Sackett DL, Spitzer WO. Methodological guidelines for systematic reviews of randomized control trials in health care from the POTSDAM consultation on meta-analysis. *J Clin Epidemiol.* 1995; 48:167-171.
31. Agency for Health Care Policy and Research. AHCPR to collaborate with AMA and AAHP to develop a national guideline clearinghouse. *Res Activ.* 1997;205:16.